APPENDIX A

SURVEY DESIGN AND CALCULATION OF NATIONAL ESTIMATES

APPENDIX A

SURVEY DESIGN AND CALCULATION OF NATIONAL ESTIMATES

In 1998, EPA distributed two industry surveys that were similar in content and purpose. The first survey, entitled U.S. EPA Collection of 1997 Iron and Steel Industry Data (detailed survey), was mailed to 176 iron and steel industry sites. The second survey, entitled U.S. EPA Collection of 1997 Iron and Steel Industry Data (Short Form) (short survey), was mailed to 223 iron and steel industry sites. Both surveys collected detailed technical and financial information from iron and steel industry sites. The short form is an abbreviated version of the detailed survey and was designed for those iron and steel sites that do not have manufacturing processes found only at integrated and non-integrated mills. Section 3 of this document describes these surveys in greater detail.

Section 1 of this appendix describes the sampling plan (identification of facilities in the industry, sample design, selection of the sample, and out-of-scope and nonresponding facilities). Section 2 of this appendix describes the calculation of sample weights. Section 3 of this appendix describes the methodology for estimating national totals and their variance estimates.

1.0 SAMPLING PLAN

This section describes the development of the sampling plan, which includes identification of the iron and steel industry, selection of the facilities to receive the detailed and short surveys, and the treatment of out-of scope and nonresponding facilities.

1.1 Sampling Frame

To produce a mailing list of facilities for the detailed survey and short form, EPA developed a sampling frame of the iron and steel industry. A sampling frame is a list of all members (sampling units) of a population, from which a random sample of members will be drawn for the survey. Therefore, a sample frame is the basis for the development of a sampling plan to select a random sample. Using the sources identified in Table A-1, EPA developed a sample frame of iron and steel facilities and divided it into 12 strata (categories) based on the types of operations conducted at the facility. A sample frame size (N) is the total number of members in the frame. Since the sample frame sufficiently covered the iron and steel population, the frame size gave a good estimate of the population size (total number of elements in the population.)

EPA cross-referenced the sources in Table A-1 with one another to obtain facility level information and to ensure the accuracy and applicability of each facility's information. After removing the duplicate entries, EPA identified 822 candidate facilities to receive surveys. These candidates include some facilities that EPA now proposes to include in the Metal Products and Machinery (MP&M) Category and will be regulated under 40 CFR Part 438.

1.2 Sample Design

To minimize the burden on the respondents to the industry surveys and improve the precision of estimates from the survey, EPA grouped the facilities into 12 strata (categories), with operations in each stratum expected to be similar. In general, the strata were determined by EPA's understanding of the manufacturing processes at each facility. This grouping of similar facilities is known as stratification. Table A-2 describes the stratification of the iron and steel industry. The Agency also developed two "certainty strata," one for the detailed survey and one for the short form (strata 5 and 8, respectively).

EPA selected a stratified random sample using the sampling frame. A stratified random sample separates the eligible population into nonoverlapping strata, that are as homogeneous as possible. Together these strata make up the whole eligible population. A simple random sample is then selected from each stratum.

For the iron and steel industry surveys, there were 12 strata: seven for the detailed survey and five for the short survey. Table A-2 includes the strata descriptions.

1.3 Sample Selection of Facilities

EPA selected 402 facilities out of the 822 facilities identified in the sample frame is sample facilities to receive surveys. Table A-2 provides the frame size and sample size for each of the 12 strata. Depending on the amount/type of information EPA determined it needed for this rulemaking and the number of facilities in a stratum, the Agency either solicited information from all facilities within a stratum (i.e., performed a census) or selected a random sample of facilities within each stratum. EPA sent a survey to all the facilities in strata 5 and 8, determining that it was necessary to capture the size, complexity, or uniqueness of the steel operations present at these sites. EPA also sent surveys to all the facilities in strata 1 though 4 (all cokemaking sites, integrated steel sites, and all sintering and direct reduced iron sites) because the number of sites is relatively low and because of the size, complexity, and uniqueness of raw material preparation and steel manufacturing operations present. EPA statistically sampled the remaining sites in strata 6, 7, and 9 through 12. The sample sizes were determined to detect a relative difference of 30 percent on a proportion of 0.25 with 90 percent confidence for a binary variable (e.g., a yes/no question)[1]. EPA used the following formula to calculate the sample size for each stratum:

$$n_h = \frac{Z^2 q/(d^2 p)}{1 + \frac{[Z^2 q/(^2 p)]}{N_h}}$$

[1] While many questions are not binary, this is a common assumption used in survey methodology.

where:

n_h	=	Number of samples to be selected from stratum h, and h=1,2,...,12;
p	=	True proportion being estimated (assuming to be 0.25);
q	=	1-p;
Z	=	Value obtained from the standard normal (Z) distribution. (For 90 percent confidence, this value is 1.645, which is 95th percentile of standard normal distribution.)
d	=	Relative difference (assuming to be 0.3 or 30 percent); and
N_h	=	Total number of facilities in stratum h.

1.4 Out-of-Scope Sites and Response Rates

EPA mailed industry surveys to all of the facilities in the sample. After receiving the industry survey, EPA determined that some facilities were "out-of-scope" or "ineligible" because the regulation would not apply to them. After reviewing the survey responses, EPA identified additional ineligible facilities. In all, EPA identified 203 of the 402 sample facilities as ineligible. Over 75 percent of these facilities were ineligible because EPA is proposing that their operations be regulated under the MP&M Category (see Section 1 of this document).

Of the remaining 199 facilities, 188 were eligible respondents, and 11 were nonrespondents (i.e., did not return a survey). The overall unweighted response rate was 94 percent (188/199). Section 2 of this appendix provides detailed facility level response rates by stratum. EPA made a nonrespondent adjustment to the weights, as described in Section 2 of this appendix.

2.0 Calculation of Sample Weights

This section describes the methodology used to calculate the base weights, non-response adjustments, and the final weights. The base weights and nonresponse adjustments reflect the probability of selection for each facility and adjustments for facility level non-responses, respectively. Weighting the data allows inferences to be made about all eligible facilities, not just those included in the sample, but also those not included in the sample or those that did not respond to the survey. Also, the weighted estimates have a smaller variance than unweighted estimates (see Section 3 of this appendix for variance estimation.) In its analysis, EPA applied sample weights to survey data.

2.1 Base Weights

The base weight assigned to each facility is the reciprocal of the probability that the facility was sampled for the particular stratum. EPA took a census for strata 1 through 5 and stratum 8; thus, the probability of selection for facilities in these strata is one. EPA selected a simple random sample from strata 6 and 7 and strata 9 through 12. The probability of selection for facility *I* from stratum *h* can be written as:

$$PROBSEL_{hi} = \frac{n_h}{N_h}$$

where:

i	=	Facility i;
h	=	Any of the *h*=1,2,..., 12 strata;
n_h	=	Total sample size for stratum *h*; and
N_h	=	Total frame size for stratum *h*.

The base weight is the inverse of this probability, and for facility *I* in stratum *h* can be written as:

$$BASE\ WEIGHT_h = \frac{1}{PROBSEL_h} = \frac{N_h}{n_h}$$

Table A-2 provides the sample size and frame size by stratum. Using stratum 6 from Table 3-1 as an example, the probability of selection for all sampled facilities in stratum 6 would be 40/69=.57971. Thus, the base weight for all facilities in stratum 6 would be 1/.57971=1.725.

2.2 Facility Level Nonresponse Adjustment

EPA made a facility-level nonresponse adjustment to account for those facilities that did not complete the industry surveys. Since the eligibility status of the nonrespondents was unknown, EPA assumed that the eligibility status of the nonrespondents was proportional to the known proportion of eligible respondents and ineligibles.

The facility-level nonresponse adjustment for stratum *h* was calculated as:

$$NRA_h = \frac{n_h}{r_h}$$

where:

r_h	=	Number of sample facilities (eligible and ineligible facilities) in stratum h responding to the detailed survey and short form.

For example, the nonresponse adjustment for stratum 6 can be calculated as follows:

$$NRA_6 = \frac{40}{30 + 9} = \frac{40}{39} = 1.02564$$

Table A-3 provides the response status of the sampled cases and the base weight and facility-level nonresponse adjustment by stratum. There were no eligible respondents in stratum 12; therefore, EPA also assumed the nonrespondents to be ineligible.

2.3 Final Weights

The final facility weight is the product of the base weight and the facility-level nonresponse adjustment. This can be written as:

$$FINALWT_h = BASEWT_h \times NRA_h$$

Again, using the example from stratum 6, the final facility weight would be:

$$1.725 \times 1.02564 = 1.76923$$

Ineligible facilities also have a base weight and nonresponse adjustments, and thus an associated final weight. However, they represent only other ineligible facilities in this sample frame. Therefore, their contribution to the national estimates are not of interest, and thus their final weights are zeros.

Table A-4 provides the base weight, facility-level nonresponse adjustment factor, and final weight for each facility by stratum.

3.0 ESTIMATION METHODOLOGY

This section presents the general methodology and equations for calculating estimates from the detailed survey and short form sampling efforts.

3.1 National Estimates

For each characteristic of interest (e.g., number of a particular operation using dry air pollution control or annual discharge flow from a particular operation), EPA estimated totals for the entire U.S. iron and steel industry ('national estimates'). Each national estimate, $\hat{Y}_{st}$, was calculated as:

$$\hat{Y}_{st} = \sum_{h=1}^{12} [FINALWT_h \cdot \sum_{i=1}^{n_h} y_{hi}]$$

where:

h	=	Stratum and *h*=1,2,...12 since there are 12 strata;
$FINALWT_h$	=	Final weight for the stratum *h*; and
y_{ih}	=	Ith value from the sample in stratum *h*.

3.2 Variance Estimation

The estimate of the variance for a national estimate can be calculated as follows:

$$\mathrm{Var}(\hat{Y}_{st}) = \sum_{h=1}^{L} \mathrm{FINALWT}_h^2 \cdot \mathrm{FPC}_h \cdot n_h \cdot s_h^2$$

where:

$\hat{Y}_{st}$ = National estimate of number of facilities with the characteristic of interest;

L = Number of strata (L= 12);

$\mathrm{FPC}_h = 1 - \frac{n_h}{N_h}$ (finite population correction for stratum h); and

$s_h^2 = \frac{1}{n_h - 1}\left[\sum_{i=1}^{n_h} (y_{ih} - \bar{y}_h)^2\right]$

(the estimate of the variance within stratum h where

$\bar{y}_h = \frac{\sum_{i=1}^{h_h} y_{ih}}{n_h}$ is the sample mean of stratum h).

The variance estimates can be used to calculate confidence intervals for the survey estimates. The confidence interval comprises a lower confidence limit and an upper confidence limit. The greater the variance, the wider the interval, and the lower the precision associated with the estimate. A 95-percent confidence interval should be interpreted as follows: If many samples were taken from the population of interest and a confidence interval were calculated from each sample, 95 percent of the confidence intervals would contain the true value of what is being estimated and 5 percent of the confidence intervals would not contain the true value. Thus, a 95-percent confidence interval is interpreted as saying that the true value of the population can be found by the random interval 95 percent of the time. The lower and upper 95-percent confidence limits can be written as:

Lower 95-percent confidence limit = $\hat{Y}_{st} - (Z_{0.025} \cdot \sqrt{\mathrm{var}(\hat{Y}_{st})})$

Upper 95-percent confidence limit = $\hat{Y}_{st} - (Z_{0.025} \cdot \sqrt{\mathrm{var}(\hat{Y}_{st})})$

where:

$Z_{0.025}$ = Value obtained from the standard normal (Z) distribution. (For 95-percent confidence interval, this value is 1.96, which is 97.5th percentile of standard normal distribution.)[2]

When comparing estimates, if the confidence intervals overlap, there is no statistically significant difference between the two estimates.

4.0 REFERENCES

A-1 Cochran, William G. *Sampling Techniques* , 3rd ed., New York: John Wiley and Sons, Inc., 1977.

A-2 *SAS*®, The SAS System, SAS Institute Inc.

[2]When the national estimate is based on a sample size of less than 30, the appropriate value from the t distribution is used instead of $Z_{0.025}$ for calculating the upper and lower confidence limits.

Table A-1

Sources Used For Development of Sample Frame

1	Association of Iron and Steel Engineers' 1997 Directory: Iron and Steel Plants Volume 1, Plants and Facilities
2	Iron and Steel Works of the World (12th edition) directory
3	Iron and Steel Society's Steel Industry of Canada, Mexico, and the United States: Plant Locations Map
4	American Coke and Coal Chemicals Institute (Membership List)
5	American Galvanizers Association (Membership List)
6	American Iron and Steel Institute (Membership List)
7	American Wire Producers Association (Membership List)
8	Cold Finished Steel Bar Institute (Membership List)
9	Specialty Steel Industry of North America (Membership List)
10	Steel Manufacturers Association (Membership List)
11	Steel Tube Industry of North America (Membership List)
12	Wire Association International (Membership List)
13	Dun & Bradstreet Facility Index database
14	EPA Permit Compliance System (PCS) database
15	EPA Toxic Release Inventory (TRI) database
16	Iron and Steelmaker Journal, "Roundup" editions
17	33 Metalproducing Journal, "Census of the North American Steel Industry"
18	33 Metalproducing Journal, "Roundup" editions

Table A-2

Frame Sizes and Sample Sizes for the Iron and Steel Population Frame

Stratum h	Stratum Description	Frame Size (N_h)	Sample Size (n_h)
Detailed Survey Strata			
1	Integrated steel facilities with cokemaking	9	9
2	Integrated steel facilities without cokemaking	12	12
3	Stand-alone cokemaking facilities	16	16
4	Stand-alone direct reduced ironmaking or sintering facilities	5	5
5	Detailed survey certainty stratum	60	60
6	Non-integrated facilities (with and without finishing)	69	40
7	Stand-alone finishing and stand-alone hot forming facilities	54	35
Short Survey Strata			
8	Short survey certainty stratum	13	13
9	Stand-alone cold forming facilities	62	37
10	Stand-alone pipe and tube facilities	164	59
11	Stand-alone hot dip coating facilities	106	49
12	Stand-alone wire facilities	252	67
	TOTAL:	**822**	**402**

Table A-3

Response Status, Base Weight, and Facility-Level Nonresponse Adjustments by Stratum

Stratum (h)	Frame Size ($N_{h)}$	Sample Size (n_h)	Response Status			Base Weight	Facility Level Nonresponse Adjustment
			Number of Eligible	Number of Ineligible	Number of Nonrespondents		
1	9	9	9	0	0	1.00000	1.00000
2	12	12	12	0	0	1.00000	1.00000
3	16	16	15	1	0	1.00000	1.00000
4	5	5	3	2	0	1.00000	1.00000
5	60	60	54	4	2	1.00000	1.03448
6	69	40	30	9	1	1.72500	1.02564
7	54	35	28	7	0	1.54286	1.00000
8	13	13	11	2	0	1.00000	1.00000
9	62	37	19	18	0	1.67568	1.00000
10	164	59	6	50	3	2.77966	1.05357
11	106	49	1	48	0	2.16327	1.00000
12	252	67	0	62	5	3.76119	0.00000
Total	**822**	**402**	**188**	**203**	**11**		

Table A-4

Base Weights, Facility-Level Nonresponse Adjustment Factors, and Final Weights by Stratum

Stratum	Base Weight	Facility Level Nonresponse Adjustment	Final Weight
1	1.00000	1.00000	1.00000
2	1.00000	1.00000	1.00000
3	1.00000	1.00000	1.00000
4	1.00000	1.00000	1.00000
5	1.00000	1.03448	1.03448
6	1.72500	1.02564	1.76923
7	1.54286	1.00000	1.54286
8	1.00000	1.00000	1.00000
9	1.67568	1.00000	1.67568
10	2.77966	1.05357	2.92857
11	2.16327	1.00000	2.16327
12	3.76119	0.00000	0.00000

APPENDIX B

MODIFIED DELTA-LOGNORMAL DISTRIBUTION

APPENDIX B
MODIFIED DELTA-LOGNORMAL DISTRIBUTION

This appendix describes the modified delta-lognormal distribution and the estimation of the episode-specific long-term averages and variability factors used to calculate the limitations and standards.[1] This appendix provides the statistical methodology that was used to obtain the results presented in Section 14.

[1]In the remainder of this appendix, references to 'limitations' includes 'standards.'

B.1 Basic Overview of the Modified Delta-Lognormal Distribution

EPA selected the modified delta-lognormal distribution to model pollutant effluent concentrations from the iron and steel industry in developing the long-term averages and variability factors. A typical effluent data set from a sampling episode or self-monitoring episode (see Section 12 for a discussion of the data associated with these episodes) consists of a mixture of measured (detected) and non-detected values. The modified delta-lognormal distribution is appropriate for such data sets because it models the data as a mixture of measurements that follow a lognormal distribution and non-detect measurements that occur with a certain probability. The model also allows for the possibility that non-detect measurements occur at multiple sample-specific detection limits. Because the data appeared to fit the modified delta-lognormal model reasonably well, EPA has determined that this model is appropriate for these data.

The modified delta-lognormal distribution is a modification of the 'delta distribution' originally developed by Aitchison and Brown.[2] While this distribution was originally developed to model economic data, other researchers have shown the application to environmental data.[3] The resulting mixed distributional model, that combines a continuous density portion with a discrete-valued spike at zero, is also known as the delta-lognormal distribution. The delta in the name refers to the proportion of the overall distribution contained in the discrete distributional spike at zero, that is, the proportion of zero amounts. The remaining non-zero, non-censored (NC) amounts are grouped together and fit to a lognormal distribution.

EPA modified this delta-lognormal distribution to incorporate multiple detection limits. In the modification of the delta portion, the single spike located at zero is replaced by a discrete distribution made up of multiple spikes. Each spike in this modification is associated with a distinct sample-specific detection limit associated with non-detected (ND) measurements in the

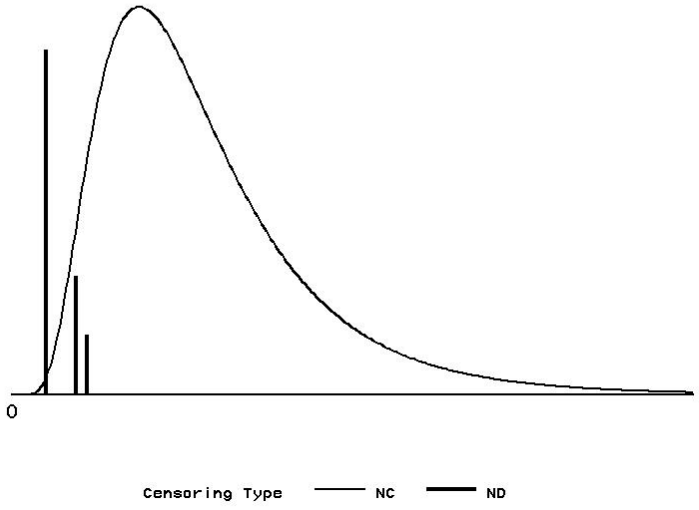
---

[2]Aitchison, J. and Brown, J.A.C. (1963) The Lognormal Distribution. Cambridge University Press, pages 87-99.

[3]Owen, W.J. and T.A. DeRouen. 1980. "Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants." *Biometrics*, 36:707-719.

database.[4] A lognormal density is used to represent the set of measured values. This modification of the delta-lognormal distribution is illustrated in Figure 1.

Figure 1
Modified Delta-Lognormal Distribution



The following two subsections describe the delta and lognormal portions of the modified delta-lognormal distribution in further detail.

B.2 Continuous and Discrete Portions of the Modified Delta-Lognormal Distribution

In the discrete portion of the modified delta-lognormal distribution, the non-detected values corresponding to the k reported sample-specific detection limits. In the model, δ represents the proportion of non-detected values and is the sum of smaller fractions, δ_i, each representing the proportion of non-detected values associated with each distinct detection limit value. By letting D_i equal the value of the i^{th} smallest distinct detection limit in the data set and the random variable X_D represents a randomly chosen non-detected measurement, the cumulative

[4]Previously, EPA had modified the delta-lognormal model to account for non-detected measurements by placing the distributional "spike" at a single positive value, usually equal to the nominal method detection limit, rather than at zero. For further details, see Kahn and Rubin, 1989. This adaptation was used in developing limitations and standards for the organic chemicals, plastics, and synthetic fibers (OCPSF) and pesticides manufacturing rulemakings. EPA has used the current modification in several, more recent, rulemakings.

distribution function of the discrete portion of the modified delta-lognormal model can be mathematically expressed as:

$$\Pr\left(X_D \le c\right) = \frac{1}{\delta} \sum_{i:D_i \le c} \delta_i \qquad 0 < c \tag{1}$$

The mean and variance of this discrete distribution can be calculated using the following formulas:

$$E\left(X_D\right) = \frac{1}{\delta} \sum_{i=1}^{k} \delta_i \, D_i \tag{2}$$

$$Var\left(X_D\right) = \frac{1}{\delta} \sum_{i=1}^{k} \delta_i \left(D_i - E\left(X_D\right)\right)^2 \tag{3}$$

The continuous, lognormal portion of the modified delta-lognormal distribution was used to model the detected measurements from the iron and steel industry database. The cumulative probability distribution of the continuous portion of the modified delta-lognormal distribution can be mathematically expressed as:

$$\Pr\left[X_C \le c\right] = \Phi\left[\frac{\ln(c) - \mu}{\sigma}\right] \tag{4}$$

where the random variable X_C represents a randomly chosen detected measurement, Φ is the standard normal distribution, and μ and σ are parameters of the distribution.

The expected value, $E(X_C)$, and the variance, $Var(X_C)$, of the lognormal distribution can be calculated as:

$$E\left(X_C\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{5}$$

$$Var\left(X_C\right) = \left[E\left(X_C\right)\right]^2 \left(\exp\left(\sigma^2\right) - 1\right) \tag{6}$$

B.3 Combining the Continuous and Discrete Portions

The continuous portion of the modified delta-lognormal distribution is combined with the discrete portion to model data sets that contain a mixture of non-detected and detected measurements. It is possible to fit a wide variety of observed effluent data sets to the modified delta-lognormal distribution. Multiple detection limits for non-detect measurements are incorporated, as are measured ("detected") values. The same basic framework can be used even if there are no non-detected values in the data set (in this case, it is the same as the lognormal distribution). Thus, the modified delta-lognormal distribution offers a large degree of flexibility in modeling effluent data.

The modified delta-lognormal random variable U can be expressed as a combination of three other independent variables, that is,

$$U = I_u X_D + (1 - I_u) X_C \tag{7}$$

where X_D represents a random non-detect from the discrete portion of the distribution, X_C represents a random detected measurement from the continuous lognormal portion, and I_u is an indicator variable signaling whether any particular random measurement, u, is non-detected or non-censored (that is, I_u=1 if u is non-detected; I_u=0 if u is non-censored). Using a weighted sum, the cumulative distribution function from the discrete portion of the distribution (equation 1) can be combined with the function from the continuous portion (equation 4) to obtain the overall cumulative probability distribution of the modified delta-lognormal distribution as follows,

$$\Pr(U \le c) = \sum_{i:D_i \le c} \delta_i + (1 - \delta)\Phi\left[\frac{\ln(c) - \mu}{\sigma}\right] \tag{8}$$

where D_i is the value of the i^{th} sample-specific detection limit.

The expected value of the random variable *U* can be derived as a weighted sum of the expected values of the discrete and continuous portions of the distribution (equations 2 and 5, respectively) as follows

$$E(U) = \delta E(X_D) + (1-\delta) E(X_C) \tag{9}$$

In a similar manner, the expected value of the random variable squared can be written as a weighted sum of the expected values of the squares of the discrete and continuous portions of the distribution as follows

$$E(U^2) = \delta E(X_D^2) + (1-\delta) E(X_C^2) \tag{10}$$

Although written in terms of U, the following relationship holds for all random variables, U, X_D, and X_C.

$$E(U^2) = Var(U) + [E(U)]^2 \tag{11}$$

So using equation 11 to solve for Var(U), and applying the relationships in equations 9 and 10, the variance of *U* can be obtained as

$$\mathrm{Var}(U) = \delta\left(\mathrm{Var}(X_D) + [\mathrm{E}(X_D)]^2\right) + (1-\delta)\left(\mathrm{Var}(X_C) + [\mathrm{E}(X_C)]^2\right) - [\mathrm{E}(U)]^2 \tag{12}$$

B.4 Autocorrelation

Effluent data from wastewater treatment technologies may be autocorrelated. For example, autocorrelation would be present in the data if the loading of a pollutant is relatively high one day, and is likely to remain high the next, and possibly, succeeding days. The measurements may be similar from one day to the next because of retention of wastewater in basins, holding ponds, and other components of the wastewater system. For data with autocorrelation, statistical time series are appropriate for modeling the data.

There are many time series models that might be considered for modeling wastewater measurements. One method of modeling autocorrelation is by using an autoregressive lag-1 model, designated as an AR(1) model. The AR(1) model is a reasonable model for many series of wastewater measurements. The AR(1) model has one parameter, ρ, the correlation between

the measurements from successive sampling events, of which time intervals are equally spaced, otherwise referred to as the lag-1 correlation. Unless specified, ρ is assumed to be zero.

The autocorrelation affects the mean and variance estimates for the data. The autocorrelation adjustments account for the effects of autocorrelation on these estimates. These adjustments are discussed in the following sections.

B.5 Episode-specific Estimates Under the Modified Delta-Lognormal Distribution

In order to use the modified delta-lognormal model to calculate the limitations, the parameters of the distribution are estimated from the data. These estimates are then used to calculate the limitations.

The parameters δ_i and δ are estimated from the data using the following formulas:

$$\hat{\delta}_i = \frac{1}{n}\sum_{j=1}^{n_d} I\left(d_j = D_i\right)$$
$$\hat{\delta} = \frac{n_d}{n} \qquad (13)$$

where n_d is the number of non-detected measurements, d_j, j = 1 to n_d, are the detection limits for the non-detected measurements, n is the number of measurements (both detected and non-detected) and I(·) is an indicator function equal to one if the phrase within the parentheses is true and zero otherwise. The "hat" over the parameters indicates that they are estimated from the data.

The expected value and the variance of the discrete portion of the modified delta-lognormal distribution can be estimated from the data as:

$$\hat{E}\left(X_D\right) = \frac{1}{\hat{\delta}}\sum_{i=1}^{k} \hat{\delta}_i D_i \qquad (14)$$

$$\hat{V}ar\left(X_D\right) = \frac{1}{\hat{\delta}} \sum_{i=1}^{k} \hat{\delta}_i \left(D_i - \hat{E}\left(X_D\right)\right)^2 \tag{15}$$

The parameters of the continuous portion of the modified delta-lognormal distribution, μ and σ^2 are estimated by

$$\hat{\mu} = \sum_{i=1}^{n_c} \frac{\ln\left(x_i\right)}{n_c}$$
$$\hat{\sigma}^2 = \frac{1}{g\left(\rho_C\right)} \sum_{i=1}^{n_c} \frac{\left(\ln\left(x_i\right) - \hat{\mu}\right)^2}{n_c - 1} \tag{16}$$

where x_i is the ith detected measurement value and n_c is the number of detected measurements (note that $n = n_d + n_c$), and $g(\rho_C)$ adjusts the estimate of σ^2 for the effects of autocorrelation to create an unbiased estimate for σ^2. The adjustment for autocorrelation is:

$$g\left(\rho_C\right) = 1 - \frac{2}{n_c\left(n_c - 1\right)} \frac{\rho_C}{1 - \rho_C} \left(n_c - 1 - \frac{\rho_C\left(1 - \rho_C{}^{n_c - 1}\right)}{1 - \rho_C} \right) \tag{17}$$

where ρ_C is the correlation of the natural logarithm of detected measurements from successive sampling events since the lognormal model is used for continuous measurements. Note that if autocorrelation is not present in the data, $g(\rho_C)=1$.

The expected value and the variance of the lognormal portion of the modified delta-lognormal distribution can be calculated from the data as:

$$\hat{E}\left(X_C\right) = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) \tag{18}$$

$$\hat{V}ar\left(X_C\right) = \left(\hat{E}\left(X_C\right)\right)^2 \left(\exp\left(\hat{\sigma}^2\right) - 1\right) \tag{19}$$

Finally, the expected value and variance of the modified delta-lognormal distribution can be estimated using the following formulas:

$$\hat{E}(U) = \hat{\delta}\hat{E}(X_D) + (1-\hat{\delta})\hat{E}(X_C) \quad (20)$$

$$\hat{Var}(U) = \hat{\delta}\left(\hat{Var}(X_D) + \left[\hat{E}(X_D)\right]^2\right) + (1-\hat{\delta})\left(\hat{Var}(X_C) + \left[\hat{E}(X_C)\right]^2\right) - \left[\hat{E}(U)\right]^2 \quad (21)$$

Equations 18 through 21 are particularly important in the estimation of episode-specific long-term averages and variability factors as described in the following sections. These sections are preceded by a section that identifies the episode data set requirements.

B.5.1 Episode Data Set Requirements

The parameter estimates for the lognormal portion of the distribution can be calculated with as few as two distinct detected values in a data set. (In order to calculate the variance of the modified delta-lognormal distribution, two distinct detected values are the minimum number that can be used and still obtain an estimate of the variance for the distribution.)

If an episode data set for a pollutant contained three or more observations with two or more distinct detected concentration values, then EPA used the modified delta-lognormal distribution to calculate long-term averages and variability factors. If the episode data set for a pollutant did not meet these requirements, EPA used an arithmetic average to calculate the episode-specific long-term average and excluded the dataset from the variability factor calculations (because the variability could not be calculated).

In statistical terms, each measurement was assumed to be identically distributed within the episode data set.

The next two sections apply the modified delta-lognormal distribution to the data for estimating episode-specific long-term averages and variability factors for the iron and steel industry.

B.5.2 Estimation of Episode-specific Long-Term Averages

If an episode dataset for a pollutant meets the requirements described in the last section, then EPA calculated the long-term average using equation 20. Otherwise, EPA calculated the long-term average as the arithmetic average of the daily values where the sample-specific detection limit was used for each non-detected measurement.

B.5.3 Estimation of Episode-Specific Variability Factors

For each episode, EPA estimated the daily variability factors by fitting a modified delta-lognormal distribution to the measurements for each pollutant. In contrast, EPA estimated monthly variability factors by fitting a modified delta-lognormal distribution to the monthly averages for the pollutant at the episode. EPA developed these averages using the same number of measurements as the assumed monitoring frequency for the pollutant. EPA is assuming that all pollutants will be monitored weekly (approximately four times a month).[5]

B.5.3.1 Estimation of Episode-specific Daily Variability Factors

The episode-specific daily variability factor is a function of the expected value and the 99th percentile of the modified delta-lognormal distribution fit to the concentration values of the pollutant in the wastewater from the episode. The expected value was estimated using equation 20 (the expected value is the same as the episode-specific long-term average).

The 99th percentile of the modified delta-lognormal distribution fit to each data set was estimated by using an iterative approach. First, the pollutant-specific detection limits were ordered from smallest to largest. Next, the cumulative distribution function, p, for each detection limit was computed. The general form, for a given value c, was:

[5]Compliance with the monthly average limitations will be required in the final rulemaking regardless of the number of samples analyzed and averaged.

$$p = \sum_{i:D_i \le c} \hat{\delta}_i + \left(1-\hat{\delta}\right)\Phi\left[\frac{\ln(c)-\hat{\mu}}{\hat{\sigma}}\right] \qquad (22)$$

where Φ is the standard normal cumulative distribution function. Next, the interval containing the 99th percentile was identified. Finally, the 99th percentile of the modified delta-lognormal distribution was estimated. The following steps were completed to compute the estimated 99th percentile of each data subset:

Step 1 Using equation 22, k values of p at $c=D_m$, m=1,...,k were computed and labeled p_m.

Step 2 The smallest value of m (m=1,...,k), such that $p_m \geq 0.99$, was determined and labeled as p_j. If no such m existed, steps 3 and 4 were skipped and step 5 was computed instead.

Step 3 Computed $p^* = p_j - \delta_j$.

Step 4 If $p^* < 0.99$, then $\hat{P}99 = D_j$

else if $p^* \geq 0.99$, then

$$\hat{P}99 = \exp\left(\hat{\mu}+\hat{\sigma}\Phi^{-1}\left[\frac{0.99-\sum_{i=1}^{j-1}\hat{\delta}_i}{1-\hat{\delta}}\right]\right) \qquad (23)$$

where Φ^{-1} is the inverse normal distribution function.

Step 5 If no such m exists such that $p_m \geq 0.99$ (m=1,...,k), then

$$\hat{P}99 = \exp\left(\hat{\mu}+\hat{\sigma}\Phi^{-1}\left[\frac{0.99-\hat{\delta}}{1-\hat{\delta}}\right]\right) \qquad (24)$$

The episode-specific daily variability factor, VF1, was then calculated as:

$$VF1 = \frac{\hat{P}99}{\hat{E}(U)} \tag{25}$$

B.5.3.2 Estimation of Episode-Specific Monthly Variability Factors Assuming No Autocorrelation

EPA estimated the monthly variability factors by fitting a modified delta-lognormal distribution to the monthly averages. Episode-specific monthly variability factors were based on 4-day monthly averages because the monitoring frequency assumed to be weekly (approximately four times a month).

In order to calculate the 4-day variability factors (VF4), the assumption was made that the approximating distribution of $\overline{U}_4$, the sample mean for a random sample of four independent concentrations, was also derived from the modified delta-lognormal distribution.[6] To obtain the expected value of the 4-day averages, equation 20 is modified for the mean of the distribution of 4-day averages:

$$\hat{E}\left(\overline{U}_4\right) = \hat{\delta}_4 \, \hat{E}\left(\overline{X}_4\right)_D + \left(1-\hat{\delta}_4\right)\hat{E}\left(\overline{X}_4\right)_C \tag{26}$$

where $\left(\overline{X}_4\right)_D$ denotes the mean of the discrete portion of the distribution of the average of four independent concentrations, (i.e., when all observations are non-detected values) and $\left(\overline{X}_4\right)_C$ denotes the mean of the continuous lognormal portion (i.e., when any observations are detected).

First, it was assumed that the detection of each measurement is independent (the measurements were also assumed to be independent; see the following section for adjustments

[6]As described in Section 14.4, when non-detected measurements are aggregated with non-censored measurements, EPA determined that the result should be considered non-censored.

for autocorrelation). Therefore, the probability of the detection of the measurements is $\delta_4 = \delta^4$. Because the measurements are assumed to be independent, the following relationships hold:

$$\begin{aligned} \hat{E}\left(\overline{U}_4\right) &= \hat{E}\left(U\right) \\ \hat{Var}\left(\overline{U}_4\right) &= \frac{\hat{Var}\left(U\right)}{4} \\ \hat{E}\left(\left(\overline{X}_4\right)_D\right) &= \hat{E}\left(X_D\right) \\ \hat{Var}\left(\left(\overline{X}_4\right)_D\right) &= \frac{\hat{Var}\left(X_D\right)}{4} \end{aligned} \tag{27}$$

Substituting into equation 27 and solving for the expected value of the continuous portion of the distribution gives:

$$\hat{E}\left(\overline{X}_4\right)_C = \frac{\hat{E}\left(U\right) - \hat{\delta}^4\,\hat{E}\left(X_D\right)}{1-\hat{\delta}^4} \tag{28}$$

Using the relationship in equation 20 for the averages of 4-day measurements and substituting terms from equation 26 and solving for the variance of the continuous portion of $\overline{U}_4$ gives:

$$\hat{Var}\left(\overline{X}_4\right)_C = \frac{\frac{\hat{Var}\left(U\right)}{4} + \left[\hat{E}\left(U\right)\right]^2 - \hat{\delta}^4\left(\frac{\hat{Var}\left(X_D\right)}{4} + \left[\hat{E}\left(X_D\right)\right]^2\right)}{1-\hat{\delta}^4} - \left[\hat{E}\left(\overline{X}_4\right)_C\right]^2 \tag{29}$$

Using equations 18 and 19 and solving for the parameters of the lognormal distribution describing the distribution of $\left(\overline{X}_4\right)_C$ gives:

$$\hat{\sigma}_4^2 = \ln\left(\frac{\hat{Var}\left(\overline{X}_4\right)_C}{\left(\hat{E}\left(\overline{X}_4\right)_C\right)^2} + 1\right) \quad \text{and} \quad \hat{\mu}_4 = \ln\left(\hat{E}\left(\overline{X}_4\right)_C\right) - \frac{\hat{\sigma}_4^2}{2} \tag{30}$$

In finding the estimated 95th percentile of the average of four observations, four non-detects, not all at the same sample-specific detection limit, can generate an average that is not necessarily equal to D_1, D_2,..., or D_k. Consequently, more than k discrete points exist in the distribution of the 4-day averages. For example, the average of four non-detects at k=2 detection limits, are at the following discrete points with the associated probabilities:

i	D_i^*	δ_i^*
1	D_1	δ_1^4
2	$(3D_1 + D_2)/4$	$4\delta_1^3\delta_2$
3	$(2D_1 + 2D_2)/4$	$6\delta_1^2\delta_2^2$
4	$(D_1 + 3D_2)/4$	$4\delta_1\delta_2^3$
5	D_2	δ_2^4

When all four observations are non-detected values, and when k distinct non-detected values exist, the multinomial distribution can be used to determine associated probabilities. That is,

$$\Pr\left[\overline{U}_4 = \frac{\sum_{i=1}^{k} u_i D_i}{4}\right] = \frac{4!}{u_1!\,u_2!\ldots u_k!}\prod_{i=1}^{k}\delta_i^{u_i} \qquad (31)$$

where u_i is the number of non-detected measurements in the data set with the D_i detection limit. The number of possible discrete points, k^*, for k=1,2,3,4, and 5 are as follows:

k	k^*
1	1
2	5
3	15
4	35
5	70

To find the estimated 95th percentile of the distribution of the average of four observations, the same basic steps as for the 99th percentile of the distribution of the observations given in section B.5.3.1, were used with the following changes:

Step 1 Change P_{99} to P_{95}, and 0.99 to 0.95.

Step 2 Change D_m to D_m^*, the weighted averages of the sample-specific detection limits.

Step 3 Change δ_i to δ_i^*.

Step 4 Change k to k^*, the number of possible discrete points based on k detection limits.

Step 5 Change the estimates of δ, μ and σ^2 to estimates of δ^4, μ_4 , and σ_4^2 respectively.

Then, using $\hat{E}\left(\overline{U}_4\right) = \hat{E}\left(U\right)$, the estimate of the episode-specific 4-day variability factor, VF4, was calculated as:

$$VF4 = \frac{\hat{P}95}{\hat{E}\left(U\right)} \tag{32}$$

B.5.3.3 Estimation of Episode-Specific Variability Factors For Monthly Averages Assuming Autocorrelation

Autocorrelation in the successive measurements affects the variance of the monthly averages. Therefore, autocorrelation must be accounted for when calculating the monthly variability factors. The calculations of the monthly variability factors when the observations are correlated assumes that the data follow the Lag-1 AR model discussed in Section B.4 and that all values are detected. Reported detection limits for non-detected measurements are treated as measured values in the continuous portion.

Assuming that all measurements are detected is equivalent to assuming that $\_/$= 0, the data have a lognormal distribution, and the equations for the continuous portion of the delta-lognormal distribution can be adapted to describe all the data. Autocorrelation has been already incorporated into the estimates of μ and σ as in equation 16 and additional adjustment to the

monthly variance $\hat{V}ar\left(\overline{U}_4\right)$ from equation 27 is required. Once the following adjustment is incorporated, the procedure described in the previous section can be used.

Using the Lag-1 AR model discussed in Section B.4 to model the effluent data, and assuming that these effluent values follow a lognormal distribution with parameters μ and σ, the variance of the monthly averages of autocorrelated values is approximated by:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\overline{U}_4\right) = \frac{\hat{\mathrm{V}}\mathrm{ar}(U)}{4}\left(1 + f_4\left(\rho_A\right)\right) \tag{33}$$

where f_4 is the factor to adjust for the autocorrelation.

In general, the f_m factor to adjust for autocorrelation can be written as:

$$f_m\left(\rho_A\right) = \frac{1}{m}\sum_{i\in S}\sum_{\substack{j\in S \\ j\neq i}} \frac{\exp\left(\rho_A{}^{|i-j|}\hat{\sigma}^2\right)-1}{\exp\left(\hat{\sigma}^2\right)-1} \tag{34}$$

where ρ_A is the correlation of the natural logarithm of measurements from successive sampling events of the same time intervals assuming all values are non-censored and S is the set of sampling events (represented by sequential numbers) on which samples for the average are taken and m is the number of sampling events in S. For a monthly average based on 4-day samples collected a week apart, the resulting formula can be simplified to:

$$f_4\left(\rho_A\right) = \frac{2}{4}\sum_{k=1}^{3}(4-k)\frac{\exp\left(\rho_A{}^k\hat{\sigma}^2\right)-1}{\exp\left(\hat{\sigma}^2\right)-1} \tag{35}$$

B.5.3.4 Evaluation of Episode-Specific Variability Factors

The parameter estimates for the lognormal portion of the distribution can be calculated with as few as two distinct measured values in a data set (in order to calculate the variance); however, these estimates can be unstable (as can estimates from larger data sets). As stated in section B.5.1, EPA used the modified delta-lognormal distribution to develop episode-specific variability factors for data sets that had three or more observations with two or more distinct measured concentration values.

To identify situations producing unexpected results, EPA reviewed all of the variability factors and compared daily to monthly variability factors. EPA used several criteria to determine if the episode-specific daily and monthly variability factors should be included in calculating the option variability factors. One criteria that EPA used was that the daily and monthly variability factors should be greater than 1.0. A variability factor less than 1.0 would result in a unexpected result where the estimated 99th percentile would be less than the long-term average. This would be an indication that the estimate of $\hat{\sigma}$ (the standard deviation in log scale) was unstable. A second criteria was that not all of the sample-specific detection limits could exceed the values of the non-censored values. All the episode-specific variability factors used for the limitations and standards met first and second criteria. A third criteria was that the daily variability factor had to be greater than the monthly variability factor. When this criteria was not met, the daily and monthly variability factors were excluded.

B.6 References

Aitchison, J. and J.A.C. Brown. 1963. *The Lognormal Distribution.* Cambridge University Press, New York.

Barakat, R. 1976. "Sums of Independent Lognormally Distributed Random Variables." *Journal Optical Society of America*, **66:** 211-216.

Cohen, A. Clifford. 1976. Progressively Censored Sampling in the Three Parameter Log-Normal Distribution. *Technometrics*, 18:99-103.

Crow, E.L. and K. Shimizu. 1988. *Lognormal Distributions: Theory and Applications*. Marcel Dekker, Inc., New York.

Kahn, H.D., and M.B. Rubin. 1989. "Use of Statistical Methods in Industrial Water Pollution Control Regulations in the United States." *Environmental Monitoring and Assessment.* Vol. 12:129-148.

Owen, W.J. and T.A. DeRouen. 1980. Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants. *Biometrics*, 36:707-719.